Statistics Can Lie but Can also Correct for Lies: Reducing Response Bias in NLAAS via Bayesian Imputation

Statistics and Its Interface Volume 6 (2013) 387–398

Statistics can lie but can also correct for lies: Reducing response bias in NLAAS via Bayesian imputation

Jingchen Liu*, Xiao-Li Meng, Chih-nan Chen and Margarita Alegria

The National Latino and Asian American Study (NLAAS) is a large scale survey of psychiatric epidemiology, the most comprehensive survey of this kind. A unique feature of NLAAS is its embedded experiment for estimating the effect of alternative orderings of interview questions. The findings from the experiment are not completely unexpected, but nevertheless alarming. Compared to the survey results from the widely used traditional ordering, the self-reported psychiatric service-use rates are often doubled or even tripled under a more sensible ordering introduced by NLAAS. These findings explain certain perplexing empirical findings in literature, but at the same time impose some grand challenges. For example, how can one assess racial disparities when different races were surveyed with different survey instruments that are now known to induce substantial differences? The project documented in this paper is part of an effort to address these questions. It creates models for imputing the original responses had the respondents under the traditional survey not taken advantage of the skip patterns to reduce interview time, which resulted in increased rates of incorrect negative responses over the course of the interview. The imputation modeling task is particularly challenging because of the complexity of the questionnaire, the small sample sizes for subgroups of interests, and the need for providing sensible imputation to whatever sub-population that a future user might be interested in studying. As a case study, we report both our findings and frustrations in our quest for dealing with these common real-life complications.

AMS 2000 subject classifications: Primary 6207, 62P10; secondary 62D99.

Keywords and phrases: Checking imputation quality, Continuation ratio model, Mental health, Multiple imputation, Probit model, Question ordering.

*Corresponding author.

1. TRIVIAL ORDERING BUT SERIOUS BIAS

1.1 A national mental health survey

The National Latino and Asian American Study (NLAAS) is a complex interview-based survey of household residents, ages 18 or older, in the non-institutionalized Latino and Asian populations of the coterminous United States. A basic task of NLAAS is to report the prevalence of psychiatric disorders and service usage. The sample consists of 2,554 Latinos, 2,095 Asians, and 215 whites. The weighted response rates were: 73.2% for the total sample, 75.5% for the Latino and 65.5% for the Asian (see [2]). Overall, there are more than 5,000 variables, measured or constructed based on raw measures available in the data set. The data were made public in July of 2007; details can be found in [2, 12] and http://www.icpsr.umich.edu/CPES/index.html.

Survey responses are known to be influenced by many factors, including the ordering of the questions. A substantial response bias induced by ordering is observed in NLAAS for the respondents' self-reported mental health and substance use services. The bias was detected because NLAAS has two sets of questionnaire designs, the traditional design and a new design, which share the same questions, but have different ordering of questions for the service use part.

Table 1 lists 13 types of mental health and substance treatment services in NLAAS. For each service, there is a "stem question" asking if the respondent ever had this service during his/her lifetime and, if yes, had he/she used services during the past 12 months. Together with the stem question, there are 5–10 follow-up questions asking more details about the self-reported service use, such as when the respondent used the service for the first time and for the last time, how many professionals he/she ever talked to, etc. The follow-up questions were obviously skipped by the interviewer if the respondent answered negatively to the stem question. This logically correct skip pattern, however, has an unintended interaction with the ordering of the questions.

1.2 A built-in experiment in the survey

The traditional service use design adopts a *sequential ordering*. After each stem question, if the response is positive, follow-up questions are asked immediately; otherwise,

Table 1. Comparing self-reported lifetime service uses

	New Design	Old Design
1. Psychiatrist	14.9%	10.4%
2. General Practitioner	17.6%	13.1%
3. Other Medical Doctor	9.2%	3.8%
4. Psychologist	13.4%	9.7%
5. Social worker	7.6%	3.4%
6. Counselor	13.2%	8.7%
7. Other Mental Health Prof	5.3%	3.2%
8. Nurse, Occupational Therapist	4.0%	2.0%
9. Religious/Spiritual Advisor	15.3%	5.9%
10. Other Healer	5.9%	1.9%
11. Hot Line	2.3%	1.2%
12. Internet Group or Chat Room	2.9%	1.1%
13. Self Help Service	5.9%	4.1%

the next stem question is asked. The traditional design also arranges the whole module of service questions after a series of diagnostic questions for identifying psychiatric disorders. Similar service-use and follow-up questions were asked within each diagnostic section. This implies that service use questions typically come thirty minutes after the interview starts, as illustrated in the left column of Figure 1 (adopted from [8]) and by then the respondents had ample opportunities to realize the unintended benefit of the skip pattern.

This sequential design has been used in common practice with a long history (e.g., [21]). That NLAAS has an embedded experiment was due to the suspicion of its investigators that respondents who are given the traditional service questionnaire design might be more likely to under-report the actual service use for (at least) two reasons. First, because the follow-up questions are asked immediately after each stem question, respondents can quickly learn from the previous service question format and under-report to avoid follow-up questions and shorten the interview. The interview for such a detailed survey tends to be very long; for NLAAS, the average interview time is about 2.6 hours. Second, the stem questions are asked after the psychiatric diagnostic questions, which themselves provide ample opportunities for learning the skip pattern. Respondents tend to react more negatively when they run out of patience, especially when experiencing memory decay.

To investigate such an ordering effect and its impact, NLAAS included an experiment: 75% of the subjects were randomly assigned to the traditional *sequential survey design* as described, while 25% were assigned to a new *parallel design*. The new design moves all the stem questions far ahead, before the diagnostic questions, but leaves additional follow-up questions after the diagnostic questions as illustrated in the right column of Figure 1. Also, stem questions in the new design come earlier than those in the traditional design. Therefore, respondents had no opportunity to learn the time-saving benefit of a "no" answer, and by the time they realized it, it is too late!

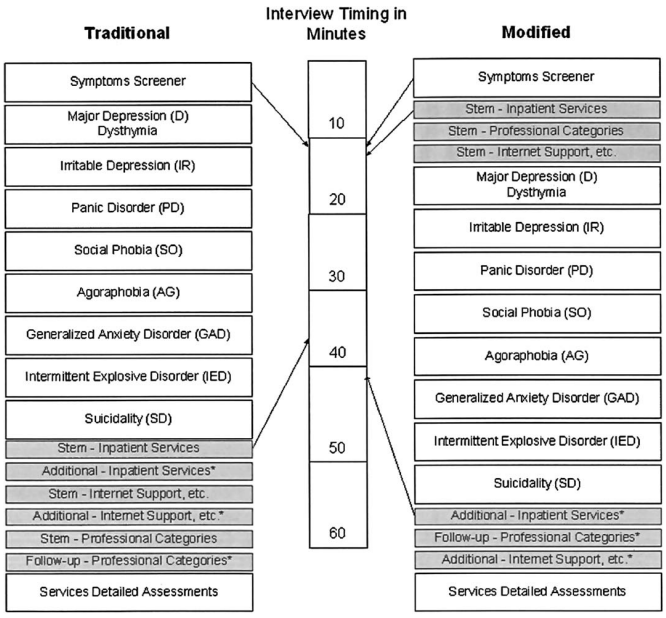

Figure 1. A chart for the first hour of the interview schedule.

The 75–25 splitting of the sample, instead of the more natural 50–50 splitting, was out of the NLAAS investigators' consideration for maintaining comparability with other collaborative surveys (for instance, the National Co-morbidity Survey Replication and the National Survey of American Life), where essentially all data were collected using the traditional sequential design. That is, if something went wrong with the new design, one would still have 75% usable data (on the service use). Unfortunately, the end results are that the sequential design is subject to serious under-reporting, as seen below. This under-reporting ultimately led to the challenging problem of correcting three quarters of data based on one quarter.

The under-reporting can be most easily seen in Table 1, which compares the (weighted) samples averages for the self-reported lifetime service use (and later in Table 3 for the past-12-month service use). The estimates from the new parallel design are uniformly higher than those from the traditional sequential design for all 13 services. The order of the questions in Table 1 is the same as they were in the actual survey. However, the last three service questions (i.e., from "Hotline" and on) were asked only once during the stem-question section, under both designs, irrespective of the type of disorders a respondent might have suffered. In contrast, the first 10 service questions were also asked *within each disorder category* (e.g., depression, panic disorder, etc.) in addition to the stem-question session. Therefore, for the first 10 questions, under the traditional design, respondents had a more repetitive task because they were asked service questions within each disorder section. This may have induced a greater number of negative responses

Table 2. Comparing other background variables (units are omitted and standard deviations are in the parenthesis)

	New Design	Old Design
Major Depression	0.120 (0.02)	0.132 (0.01)
Any Affective Disorder	0.122 (0.02)	0.136 (0.01)
Any Disorder	0.143 (0.02)	0.137 (0.01)
Any Affective Disorder 12 month	0.066 (0.01)	0.068 (0.008)
Number of disorders	0.43 (0.06)	0.43 (0.04)
k10 distress	13.75 (0.3)	13.75 (0.17)
Proportion of female	0.53 (0.03)	0.55 (0.02)
Age	41.01 (0.8)	41.05 (0.5)
Social Status*	5.57 (0.1)	5.68 (0.06)
Proportion of immigrants	0.67 (0.024)	0.68 (0.014)

*Social status: an ordinal variable taking integers from 0 to 10 indicating the relative social status.

for questions regarding services 1–10 than for services 11–13. This also explains why the under-reporting occurred even for the first service use, and that there is no significant increase in the degree of under-reporting along the ordered list (and obviously we do not expect to see a decrease either).

For the new design, the stem questions on service use were first asked before all the psychiatric disorder questions and then followed up later in each disorder section. A respondent is classified to have used a particular service, say, psychiatrist services, if the respondent reported positively to the psychiatric service question under at least one of the disorder categories or to the stem service question. Thus, learning of skip patterns during the diagnostic section would have little effect on the self-reported service use because it can only take place after the completion of all stem questions.

The phenomena of under-reporting by the traditional design persisted in sub-populations by ethnicity, as studied in detail in [8]. However, non-service variables show no significant difference between the two design groups at all. Table 2 demonstrates this for a group of randomly selected variables. It is therefore logical to conclude that the significant discrepancy in the self-reported service use rates (with p-value <0.01, which is robust to different model assumptions, as discussed in [8]), is a direct result of the different orderings of the questions. This under-reporting due to question ordering adds another example to the large literature on the impact of survey instruments on survey results (e.g. [23]).

1.3 The imputation task and the challenges

Because NLAAS serves as a public use data set, response bias can and will lead to misleading results for most potential analyses involving service use. One strategy to deal with such a problem is to create multiple imputations for the unobserved responses of those given the traditional design had they been given the new parallel design. Multiple imputation is a handy tool for dealing with incomplete data via complete data procedures; see [22, 10], and [9]. Further results on creating and analyzing multiply imputed data sets are in [4, 18, 17], and [13]. As demonstrated in these literature, Bayesian prediction is a principled approach for multiple imputation but to yield sensible results, the modeling and the associated computational tasks are often very challenging.

The imputation task would be trivial—and in fact meaningless—if our goal is just to "fix" the overall service use rates for the traditional sequential design group. A simple Bernoulli model would do the job. The ideal goal here, however, is to adjust/correct the rate for *any sub-population* that might be of interest to a potential analyst of NLAAS. This turns out to be an exceedingly difficult, indeed impossible task for NLAAS (or any similar survey), because NLAAS has more than 5,000 variables but only 4,864 subjects. In principle we should use all variables for reasons discussed in [17], but this was infeasible due to the complexity of the survey, the limitation of data, and our lack of resources. Therefore, we have to compromise by using only a set of predictive variables that are noticeably correlated with the service use *and* those that are judged to be used frequently in subsequent analysis involving service use. Assembling such a list of variables is a difficult task in itself. It is a long iterative process, based on statistical analysis and discussions with researchers in the substantive fields, and considerations of model identifiability and computational constraints.

In addition, to incorporate these variables in imputation together with the dependence among the 13 service uses, we need a suitable model for multivariate categorical data. We initially chose the multivariate probit model because of its computational and interpretational simplicity. However, we later found that the multivariate probit model is inadequate for the NLAAS data, because it is incapable of modeling high-order interactions, resulting in significant over-imputation for the combined rates (i.e., the last three rows in Table 4). Here over-imputation refers to the fact that the imputed service rates are consistently higher than the observed rates from the new design; see [15, 14] for demonstration and other preliminary imputation results. We therefore need an extension of the multivariate probit model to accommodate a higher order of interactions. This model needs to be friendly to posterior sampling because of the constraints we faced (e.g., many very time consuming test runs were necessary due to revisions of database, modifications of variables, refinements of priors, etc).

The rest of this paper summarizes our effort in these regards. Specifically, Section 2 documents our basic model assumptions including prior specifications. Section 3 discusses imputation results, investigates the issue of assessing the quality of the imputations, and concludes briefly. Due to space limitation, technical and computational details are deferred to an on-line supplement http://www.intlpress.com/SII/p/2013/6-3/SII-6-3-liu-supplement.pdf, as are materials on examining imputation quality for sub-populations.

2. CONSTRUCTING IMPUTATION MODEL

2.1 Modeling response behavior

Under a setting described in Section 1, our basic assumptions are (excluding negligible exceptions)

- *Assumption 1*: respondents who received the new parallel design responded honestly;
- *Assumption 2*: respondents who received the traditional sequential design responded honestly, if they indeed did not have service use.

These two assumptions imply that we need to impute the negative responses collected using the traditional design. One could of course ask how do we know that respondents from the new parallel design were not over-reporting? Strictly speaking, we don't, and there is no information in the data to verify Assumption 1. However, there are no compelling reasons or conceivable incentives for over-reporting under the new parallel design, unlike the strong incentive for under-reporting under the traditional sequential design for reasons listed in Section 1.2. Furthermore, even for those who disagree, our imputations can be viewed as predicting service use rates for those in the traditional design groups had they been given the new design, without referencing which group had responded correctly.

The rationale behind Assumption 2 is also common sense. Under the traditional design, there is no incentive to knowingly provide a false positive response, because the false positive can only prolong the interview. In addition, falsifying a positive response to a stem question will require the respondent to provide answers to an array of follow-up questions, a non-trivial task without actual experience.

Given these two assumptions, our basic sampling model is as follows. For each respondent, we adopt the following notations. Let y be the self-reported service use status, 1 for having service and 0 otherwise; S be the true service use status, 1 for having service and 0 otherwise; ξ be the response behavior of those people from the traditional design group who have service use (i.e., $S = 1$), 1 for responding honestly and 0 otherwise; and lastly I be the group design indicator, 1 for traditional design and 0 for new design.

The two assumptions yield that $y = S$ when $I = 0$ and $y = S\xi$ when $I = 1$. Under the new design ($I = 0$), we do not have information of ξ and thus we treat it as completely missing. For simplicity we can assume S and ξ are independent Bernoulli random variables.

2.2 A continuation ratio probit model

Recall that there are 13 services under consideration. We use subscript j to indicate different services, that is, S_j is the j-th service and ξ_j is the corresponding response behavior. Our modelling strategy for the dependence among services was motivated by the hierarchical structure of mental health and substance use services. The individual services belong to some general *types* of services, for example, the 13 services are typically grouped into 4 types: specialist, generalist, human services, and alternative services, as shown in Table 4. It is therefore reasonable to postulate "two-stage" indicators for using a particular service, that is, to use a specific service, say, psychologist, one has to first be in the category of seeing a "Specialist" and then choose or be assigned to seeing a psychologist.

Specifically, suppose the 13 services can be categorized into K types. Let $S^{(k)}$ be the indicator for the kth type, $k = 1, \ldots, K$. Within each type, suppose there are J_k services. Then, for the j-th service belonging to the k-th type, we can express our service indicator as

$$S_{k,j} = S^{(k)} \cdot S_j^{(k)}, \tag{1}$$

where $S_j^{(k)}$ is the j-th service within the k-th type, and $\{S_j^{(k)}, j = 1, \ldots, J_k\}$ and $S^{(k)}$ are assumed to be independent. We note that the expression in (1) is a special case of the continuation ratio (CR) model formulation, where the probability of the binary outcome is modeled as a product of a sequence of binary probabilities (hence continuation ratio); this is a common strategy in modeling censored survival data, see, for example, [16, 7, 20, 1, 11, 5, 6, 19].

To introduce further model flexibility, we dropped the restriction on sharing the same-type indicator and adopt a more general product form by letting

$$S_j = S_{a,j} \cdot S_{b,j}, \tag{2}$$

where $\{S_{a,j}, j = 1, \ldots, 13\}$ and $\{S_{b,j}, j = 1, \ldots, 13\}$ are mutually independent. This relaxation increases the model flexibility for achieving better fit to the data, but at the expense of its interpretability and the potential issues of over-fitting and non-identifiability. As for the interpretation, it is less appealing than that of (1), but nevertheless one can consider (2) as an attempt to let the data decide which "type" a service belongs to (e.g., the grouping in Table 4 is not written in stone; for example, it is not clear if "Hotline" should be grouped with the "Specialists" or "Human Services" or even "Alternative Service"). That is, if we view one of the two S's as a "type" indicator, then we can imagine that a posterior inference might indicate a block correlation structure where a subgroup of S's are highly correlated with each other, indicating that these services tend to be categorized together. Indeed, in the extreme case where all indicators in the same "group" are perfectly correlated, then we are back to the special case of (1).

The above interpretation would make sense, however, only when we have ways to identify which of the two S's on the right-hand-side of (2) can be associated with type and which with more specific individual services. To deal with this issue, we assume that the distribution of the type-indicator S_a is common to all respondents. This of course is a very strong assumption, but it is necessary in order to ensure identifiability because otherwise the dependence of

S_a on the covariates is generally indistinguishable, based on our observed data, from the dependence on covariates of the conditional probability of using a specific service within each type.

2.3 The likelihood specification

Our modeling process also need to take into account covariates, survey design and weights, prior specification, etc, as shall be detailed shortly. The list of covariates included in our imputation model is as follows. A set of categorical variables includes marital status, insurance status, working status, region of residency in the country, ethnicity, immigration status, gender, psychiatric disorders including any depressive disorder (lifetime and last year), any substance disorder (lifetime and last year), any anxiety disorder (lifetime and last year), and any psychiatric disorder (last year); the list of continuous variables (or can be treated as such) includes logarithm of annual income, number of psychiatric disorders, social status, age, k10 distress (a psychiatric symptom measure), and logarithm of survey weights.

With this set of covariates, we set up two independent multivariate probit models for S_a and S_b respectively. More precisely, we specify their distributions via a data augmentation scheme and associate each of S_a and S_b with a 13-dimensional multivariate normal vector (denoted by Z_a and Z_b respectively) by letting

$$S_{\zeta,j} = \mathbf{1}(Z_{\zeta,j} \geq 0), \text{ for all } \zeta = a, b \text{ and } j = 1, \ldots, 13,$$

where $\mathbf{1}(A)$ is the indicator function. The latent vectors Z_a and Z_b are assumed to be independent, and have the following distributions:

$$Z_a \sim N(\mu_a, \Sigma_a), \qquad Z_b \sim N(\beta_b^\top X + W_c, \Sigma_b), \tag{3}$$

where both Σ_a and Σ_b are covariance matrices, c indicates the survey design cluster to which each individual belongs, X is the $k \times 1$ covariate vector, and β_b is the $k \times 13$ regression parameter matrix. As it is well-known for probit models, the diagonal elements of Σ_a and Σ_b are not identifiable from the data, for which we will impose proper prior distributions. Here k is larger than the actual number of variables, denoted by $\tilde{k}$, in the model, because each categorical variable requires multiple dummy variables to represent its levels; in the current setting, we have $\tilde{k} = 19$ and $k = 39$. The clustering variable $W_c = (W_{c,1}, \ldots, W_{c,13})^\top$ is assumed to have independent normal components: $W_{c,j} \sim N(0, \alpha_j^2)$, $j = 1, \ldots, 13$ and for all c's. The use of W helps to model the cluster effect due to survey design, by allowing respondents in the survey design cluster c to share the same W_c.

For the response behavior indicators, ξ's, we fit a standard multivariate probit model with clustering (the same as S_b). We let $\xi_j = \mathbf{1}\{Z_{l,j} > 0\}$, where l is for "lie" and

$$Z_l \sim N(\beta_l^\top X + \tilde{W}_c, \Sigma_l), \tag{4}$$

with $\tilde{W}_c = (\tilde{W}_{c,1}, \ldots, \tilde{W}_{c,13})^\top$, and $\tilde{W}_{c,j} \sim N(0, \tilde{\alpha}_j^2)$.

Thererfore, for a single-subject response vector $y = (y_1, ..., y_{13})$, the likelihood can be precisely written down as for $I = 1$

$$\begin{aligned} & f(y|\mu_a, \beta_b, \Sigma_a, \Sigma_b) \\ = \; & E\Big\{ \prod_{j=1}^{13} [\mathbf{1}(Z_{a,j} \geq 0, Z_{b,j} \geq 0, Z_{l,j} \geq 0)]^{y_j} \\ & \qquad \times [\mathbf{1}(Z_{a,j} < 0, \; or \; Z_{b,j} < 0, \; or \; Z_{l,j} < 0)]^{1-y_j} \Big\}, \end{aligned}$$

for $I = 0$

$$\begin{aligned} & f(y|\mu_a, \beta_b, \Sigma_a, \Sigma_b) \\ = \; & E\Big\{ \prod_{j=1}^{13} [\mathbf{1}(Z_{a,j} \geq 0, Z_{b,j} \geq 0)]^{y_j} \\ & \qquad \times [\mathbf{1}(Z_{a,j} < 0, \; or \; Z_{b,j} < 0)]^{1-y_j} \Big\}, \end{aligned}$$

where the expatiation is taken with respect to Z_a, Z_b, and Z_l whose distributions are given by (3) and (4). In addition, we adopt the representation $0^0 = 1$.

Furthermore, exploratory data analysis shows that the observed data are not homogeneous across different ethnicity groups, especially for the dependence structure among the 13 service variables as well as among the response behaviors. However, allowing a full interaction between ethnicity and all other variables turns out to be impractical because of the sample size and computational cost. Therefore we stratify the total sample into three relatively homogenous groups: (A) Latino (including Puerto Rican, Cuban, Mexican, and other Latinos) and white people; (B) Filipino and Vietnamese; and (C) Chinese and other Asian.

We then fit a separate model within each group, including separate prior specification. The grouping here is largely determined by the response behavior. For example, as shown in Section 3, for Filipino and Vietnamese, the differences between the reported rates under the old (traditional) and new designs are much more striking than that for Chinese; see especially the last three rows in Tables 5 and 7. We do not know, however, if this similarity is because Chinese under-report substantially less on average, or they tend to under-report regardless of the design.

2.4 Prior specifications

Because we have the new design sample to match, it is possible for us to tune our prior to provide better imputations. In many cases, seeking priors to fit the model leads to over-fitting for the purpose of parameter estimation, as well as underestimating posterior uncertainty. In our situation, the goal is imputation/prediction, for which over-fitting *can be* lesser a problem. For example, the dominating part of a predictive variance is the sampling variance, which is generally not affected by over-fitting, in contrast to the posterior

variance of the parameter, which can be seriously affected. Furthermore, the model class we adopt, although more flexible than multivariate probit model, is still a relatively parsimonious class, and therefore the impact of "data snooping" is limited. The general approach we take here is to treat prior specification as an integrated part of model specification, and therefore tuning a prior is for the same purposes as improving the overall model for better prediction. (In Section 3, we will discuss in detail the issue of checking imputation quality.) The priors reported below are the ones we ended up using to produce the results given in Section 3.

To specify a prior for $(\mu_a, \beta_b, \beta_l, \Sigma_a, \Sigma_b, \Sigma_l, \alpha, \tilde{\alpha})$, we first assume that they are *a priori* independent. For the regression coefficients, β's, a constant prior will lead to an improper posterior for the response behavior coefficient β_l because the response behavior indicator is largely a latent variable. Thus, we adopt a proper prior distribution on β_l such that the under-reporting probabilities are approximately uniform. Note that it is impossible to have them all strictly uniform due to the variation of covariates among individual observations. To proceed, we first assume that all the continuous covariates in X have been standardized (across the sample $i = 1, \ldots, n$) to have sample mean 0 and sample variance 1. For any dummy variable, it is standardized by a scalar multiplier such that the sum of squares across the entire sample is n, the sample size. These standardizations of X are for convenience and will not alter the nature of the model.

We impose a relatively strong prior, $\mu_a \sim N(2, I_{13}/200)$. The choice of the mean "2" is guided by our desire that the "type" indicator S_a should not be too far from that for the standard probit model, which is equivalent to setting $S_a \equiv 1$ in (2). In any case, the significant part of this model is the introduction of Σ_a, providing flexibility for high-order interactions. For each $\beta_{l,j}$, the coefficients of service j, we choose a multivariate normal prior distribution with covariance $\tilde{\Sigma} = n(XX^\top)^{-1}/\tilde{k}$. The prior mean for $\beta_{l,j} \sim N(\mu_l, \tilde{\Sigma})$ is chosen as the following: $\mu_l^\top = (0.2, 0, \ldots, 0)$ for Group A, $\mu_l^\top = 0$ for Group B, and $\mu_l^\top = (1, 0, \ldots, 0)$ for Group C. For the prior distribution of β_b, we adopt a similar approach, and with the prior mean for $\beta_{b,j} \sim N(\mu_b, \tilde{\Sigma})$ specified as $\mu_b^\top = (-0.8, 0, \ldots, 0)$ for Group A, and $\mu_b^\top = (-1.5, 0, \ldots, 0)$ for Group B and Group C.

The prior distribution for α_j and $\tilde{\alpha}_j$ is set to be $N(0, 1)$. Note that the signs of α_j and $\tilde{\alpha}_j$ are not identifiable. But this does not affect our imputation because only α_j^2 and $\tilde{\alpha}_j^2$ enter the model. It is for computational convenience and speed that we let α live on R^1, as discussed in [24].

The prior for Σ_a, Σ_b, Σ_l is the inverse Wishart distribution ([3, 9]). In particular, we let

$$
\begin{aligned}
p(\Sigma_a) &\sim \text{Inv-Wish}(\bar{\Sigma}_a, df_a);\\
p(\Sigma_b) &\sim \text{Inv-Wish}(\bar{\Sigma}_b, df_b);\\
p(\Sigma_l) &\sim \text{Inv-Wish}(\bar{\Sigma}_l, df_l).
\end{aligned} \tag{5}
$$

Table 3. Comparing last year service use: the rate is the percentage of people reported having service last year among those who reported having service use in lifetime

	New Design	Old Design
Psychiatrist	32.9%	22.9%
Other Medical Doctor	59.5%	31.4%
Psychologist	28.6%	17.4%
Social worker	43.7%	17.6%
Counselor	21.9%	21.3%
Other Mental Health Prof	50.6%	45.0%
Nurse, Occupational Therapist	56.1%	17.8%
Religious/Spiritual Advisor	41.6%	29.8%
Hot Line	19.3%	18.8 %
Other Healer	53.8%	41.1%
Internet Group or Chat Room	64.0%	31.8%
Self Help Service	22.7%	26.9 %

We choose $\bar{\Sigma}_a = 100 I_{13}$ and $df_a = 100$. We also choose $\bar{\Sigma}_b = 16 I_{13}$, $df_b = 16$, $\bar{\Sigma}_l = 500 I_{13}$, $df_l = 500$ for Group A; $\bar{\Sigma}_b = 50 I_{13}$, $df_b = 50$, $\bar{\Sigma}_l = 500 I_{13}$, $df_l = 500$ for Group B; and $\bar{\Sigma}_b = 100(0.1 \times I_{13} + 0.9 \times \mathbf{1}\mathbf{1}^\top)$, $df_b = 100$, $\bar{\Sigma}_l = 500 I_{13}$, $df_l = 500$ for Group C.

2.5 Model for last-12-month service use

In addition to the lifetime services, NLAAS also collected data on the last-12-month service, for which similar under-reporting is observed for the traditional-design group, as shown in Table 3. When setting up the last-12-month service model, we need to obviously respect the logic constraint that is respected in the observed data: no lifetime service, not last-12-month service. Therefore, for each service, we have a bivariate random variable (S, S_T) that can take only three possible values: $\{(1,1), (1,0), (0,0)\}$, where S_T is the true last-12-month service use. This bivariate variable can be modeled as $(S, S_T) = (S, S\tilde{S})$, where S and $\tilde{S}$ are independent Bernoulli random variables, with $\tilde{S}$ being the indicator for the true last-12-month service use given the lifetime service use $S = 1$. In other words, our joint model for the self-reported lifetime use y and the past 12-month use y_T will be formulated in two stages; first the marginal model for the lifetime use and then the conditional model of the past 12-month given lifetime use.

Under our most basic model assumptions as listed in the beginning of Section 2.1, for the traditional-design group, if the observed service use is $(y, y_T) = (1, 1)$, then $(S, S_T) = (1, 1)$. If $(y, y_T) = (0, 0)$, then $\tilde{S}$ is missing. When $(y, y_T) = (1, 0)$, for which the respondent was asked about her/his last-12-month service use, s/he may really not have service use in the last 12 months or choose to under-report. In the former case, we introduce a new respondence behavior indicator for the last-12-month service use, $\tilde{\xi}$, such that it is independent of $(S, \xi, \tilde{S})$ and the reported last-12-month service use is expressed as $y_T = S\xi\tilde{S}\tilde{\xi}$. That is, $\tilde{\xi}$ is a direct analog to ξ in the lifetime service use model.

Table 4. Percentage rates of Latino lifetime service use.
New*: observed new design rates.* Imp*: imputed old design rates.* Old*: observed old design rates*

	Puerto Rican			Cuban			Mexican			Other Latino		
	New	Imp	Old	New	Imp	Old	New	Imp	Old	New	Imp	Old
Specialist (1,4,7,11)	34.2	32.3	25.5	26.7	22.8	18.0	14.9	14.8	10.2	20.8	19.4	13.9
1. Psychiatrist	28.8	22.9	16.0	19.7	16.6	12.9	11.0	9.4	6.5	9.4	11.0	8.0
4. Psychologist	20.6	18.6	14.0	16.2	13.3	9.7	10.4	8.4	5.6	13.5	12.7	8.7
7. Other M. H. Prof.	10.7	7.1	5.5	4.7	3.6	2.3	5.0	4.0	2.7	4.5	4.6	3.0
11. Hotline	1.7	2.3	1.5	2.3	0.9	0.2	0.9	1.9	1.3	3.0	2.3	1.3
Generalist (2,3,8)	32.5	30.0	25.4	21.4	23.9	19.6	18.3	16.8	12.3	16.8	15.4	10.3
2. General Practitioner	28.5	27.1	23.9	18.5	19.7	16.7	16.6	14.1	10.3	12.7	11.3	8.4
3. Other Med. Doctors	15.5	12.7	7.5	10.3	9.7	4.9	7.0	5.1	1.9	10.2	8.4	4.3
8. Other Professionals	13.4	5.6	3.5	4.0	4.2	3.2	3.0	2.7	1.7	3.3	2.4	1.2
Human Services (5,6,9)	38.5	30.5	22.7	17.4	13.6	8.1	20.1	16.6	10.2	24.9	21.1	13.3
5. Social Worker	17.9	14.0	10.1	5.1	3.3	2.6	7.4	4.9	2.6	5.5	6.0	3.8
6. Counselor	29.6	17.9	12.9	11.0	5.6	3.5	11.2	9.2	7.1	13.7	12.6	9.5
9. Religious Advisor	16.3	17.2	10.4	11.8	10.6	5.2	14.0	10.7	5.2	16.5	12.3	5.0
Alt. Services (10,12,13)	24.9	15.3	8.9	10.2	9.0	5.3	9.1	7.8	3.7	9.4	12.8	6.1
10. Other Services	13.0	8.4	1.6	6.6	5.7	2.7	4.0	3.2	1.1	4.0	6.8	2.3
12. Internet	1.8	4.4	2.5	3.2	2.1	0.6	1.4	1.6	0.1	3.7	3.3	0.6
13. Self Service	13.0	7.4	6.3	2.3	3.7	3.2	5.4	4.6	3.2	4.8	5.8	4.5
Formal Services (1–8)	46.3	47.2	40.7	33.9	35.5	29.9	25.0	26.9	20.8	29.5	31.6	25.0
Any Services (1–10)	50.0	50.4	42.4	37.6	37.6	30.6	30.0	30.1	21.9	35.3	35.9	26.2
Any Services (1–13)	50.9	51.4	42.9	38.4	38.3	31.0	32.5	31.5	22.7	36.5	37.4	26.8

Furthermore, since the likelihood for (S, ξ) and $(\tilde{S}, \tilde{\xi})$ factors, they are *a posteriori* independent under independent priors. This simplicity allows us to use independent Markov chains to sample from the posterior distributions and thereby reduce computational burden. The conditional model for the last-12-month service use is a complete analogue of the lifetime model, except for that we use multi-probit model for $\tilde{S}$ instead of the more flexible CR-probit model, because the latter does not provide sufficient improvement to outweigh its computational disadvantage.

3. IMPUTATION RESULTS AND THEIR QUALITY CHECKING

3.1 A summary of the imputation results

Our imputations were created by samples from the posterior distribution specified in the previous section. The main computational tool is Markov chain Monte Carlo (MCMC). In particular, ten imputed data sets were created by ten separate Markov chains. The detailed computational scheme is reported in Section A in the supplemental material.

Table 4 summarizes the imputation results for the lifetime service use of the Latino samples. In particular, we stratify the Latino cohort into *Puerto Rican*, *Cuban*, *Mexican*, and *Other Latino*. For each group, three columns of service rates are provided. The "New" and "Old" are the observed rates from the samples assigned respectively to the new design and traditional design; and the "Imp" is the average of the 10 imputations. With similar notation, Table 5 gives the results for the Asian population. Tables 6 and 7 are the corresponding results for the last-12-month service use.

Clearly, we cannot expect the "Imp" and "New" columns to be identical; minimally there are random variations in the covariates between the new and old groups. On the other hand, a large difference between the "Imp" and the "New" would indicate that something is amiss, for example, as with a number of service use rates for the Vietnamese group. However, determining how close is acceptable turns out to be a challenging task, as detailed in the next few sections. Here we note that, by visual inspection, the quality of our imputations appears to be better for the Latino groups than for the Asian groups. We believe this is largely due to the fact that Latino groups are more homogenous in their response behaviors, allowing us to fit them as one group and hence with more stable results due to larger sample sizes. However, even for the Latino groups, some of the imputation results are visually unsatisfactory (e.g., "Other professionals" and "Counselor" service for Puerto Rican).

Before we discuss the thorny issue of checking imputation quality, we need to address the question of the very purpose of imputation. Since essentially all information for building the imputation model comes from the new group with 25% of the total sample, one may ask why not just use the same 25% sample for subsequent analysis. Besides the logistically and politically unacceptable practice of throwing away 75% of the data, statistically, the loss of information can be much less than 75% depending on how strongly the service uses probability are determined by the fully observed

Table 5. Percentage rates of Asian lifetime service use.
New: *observed new design rates.* Imp: *imputed old design rates.* Old: *observed old design rates*

	Vietnamese			Filipino			Chinese			Other Asian		
	New	Imp	Old	New	Imp	Old	New	Imp	Old	New	Imp	Old
Specialist (1,4,7,11)	6.2	10.2	6.1	10.0	13.4	9.2	10.8	12.3	9.9	16.6	14.0	11.1
1. Psychiatrist	6.2	7.5	5.1	7.3	7.8	6.0	7.2	4.5	3.6	12.9	10.2	7.3
4. Psychologist	3.7	2.4	0.8	7.5	7.1	4.2	7.6	8.5	6.6	8.5	7.2	5.9
7. Other M. H. Prof.	3.0	2.6	1.0	4.9	4.0	1.6	4.7	1.5	0.8	2.4	2.6	1.4
11. Hotline	0.0	1.7	0.6	1.8	2.2	0.9	1.3	1.7	1.3	1.6	1.4	1.0
Generalist (2,3,8)	20.4	13.3	5.4	24.4	21.1	10.0	10.0	10.0	6.9	12.3	13.5	10.4
2. General Practitioner	18.7	10.2	5.2	21.5	16.4	8.9	10.0	9.4	6.5	11.6	11.8	9.2
3. Other Med. Doctors	8.0	5.0	1.2	12.7	8.8	3.0	5.7	1.3	0.8	2.5	4.3	2.9
8. Other Professionals	3.6	2.5	1.3	5.2	3.9	2.3	3.0	1.3	0.8	1.4	2.4	1.2
Human Services (5,6,9)	8.0	7.9	4.2	17.1	17.3	10.5	13.0	10.4	6.9	15.3	14.9	12.1
5. Social Worker	2.5	2.2	1.0	4.9	5.6	3.3	5.5	2.1	1.0	4.2	2.9	2.2
6. Counselor	3.9	4.5	2.5	13.9	10.4	6.2	6.6	6.7	5.1	8.8	9.8	7.9
9. Religious Advisor	2.7	3.8	2.3	8.0	7.2	4.4	8.2	4.9	2.7	11.1	9.7	6.1
Alt. Services (10,12,13)	6.3	5.4	2.0	9.0	9.8	3.5	8.4	6.4	4.3	7.7	10.0	6.3
10. Other Services	5.8	3.6	1.6	5.0	3.4	0.9	3.4	2.3	1.0	7.2	5.6	2.5
12. Internet	0.5	1.7	0.7	4.8	3.0	0.7	3.0	2.0	1.3	1.4	2.9	1.6
13. Self Service	0.0	1.5	0.6	5.2	5.2	2.5	2.0	2.6	2.0	1.4	4.0	3.6
Formal Services (1–8)	24.6	19.7	9.1	31.3	31.3	18.2	13.5	18.9	15.7	23.4	22.9	19.5
Any Services (1–10)	29.1	21.5	9.6	33.2	33.5	19.5	16.0	21.2	17.2	26.2	25.3	21.5
Any Services (1–13)	29.6	22.5	9.9	34.4	35.2	20.4	19.6	22.7	18.9	26.2	25.8	21.8

Table 6. Percentage rates of Latino last year service use.
New: *observed new design rates.* Imp: *imputed old design rates.* Old: *observed old design rates*

	Puerto Rican			Cuban			Mexican			Other Latino		
	New	Imp	Old	New	Imp	Old	New	Imp	Old	New	Imp	Old
Specialist (1,4,7,11)	11.4	12.6	7.8	5.7	7.9	5.1	5.2	6.1	3.1	7.6	7.1	3.3
1. Psychiatrist	10.0	8.6	5.0	5.1	6.4	4.6	3.3	4.2	1.8	2.7	4.5	1.7
4. Psychologist	3.9	5.1	2.4	2.7	3.4	1.3	2.8	3.0	1.5	3.6	4.1	2.0
7. Other M.H. Prof.	3.4	2.9	1.5	1.0	1.4	0.5	2.2	2.3	1.3	2.4	2.3	1.2
11. Hotline	0.7	0.5	0.0	0.0	0.2	0.0	0.0	0.7	0.4	0.7	0.5	0.2
Generalist (2,3,8)	21.3	19.7	6.2	14.2	15.5	5.8	9.2	9.4	3.6	10.8	8.9	3.6
2,3. Other Med. Doc.	21.3	19.3	5.9	13.3	15.4	5.6	9.2	9.2	3.5	10.8	8.3	3.5
8. Other Professionals.	4.8	2.0	0.3	2.3	1.8	0.3	0.8	1.3	0.3	1.8	1.4	0.1
Human Services (5,6,9)	16.5	13.1	4.6	7.0	5.3	0.8	8.5	8.6	2.8	14.0	9.7	3.2
5. Social Worker	6.2	4.9	1.7	3.4	1.3	0.3	2.6	2.4	0.8	2.7	1.9	0.2
6. Counselor	6.8	5.9	3.1	4.8	1.8	0.3	2.5	2.8	1.5	3.8	4.4	2.0
9. Religious Advisor	8.6	6.9	1.6	3.5	3.7	0.7	5.8	6.1	1.4	8.8	6.3	1.9
Alt Services (10,12.13)	11.5	8.1	3.3	2.5	3.5	1.1	3.4	3.8	1.4	6.8	6.6	2.3
10. Other Services	7.9	4.3	0.4	2.5	2.1	0.7	1.9	1.9	0.4	3.7	3.7	1.4
12. Internet	1.1	2.7	1.5	1.2	0.9	0.2	1.2	0.7	0.0	1.4	1.6	0.0
13. Self Services	2.6	2.5	1.6	0.0	1.0	0.6	0.7	1.9	1.1	3.4	2.8	1.2
Formal Services (1–8)	25.9	27.0	14.0	18.0	18.5	8.2	12.7	13.9	6.5	15.4	13.8	6.5
Any Services (1–10)	30.8	30.7	14.9	18.6	20.4	8.7	16.6	17.2	7.1	22.3	17.9	7.1
Any Services (1–13)	31.9	31.8	16.2	18.6	21.2	9.2	17.7	17.9	7.5	23.0	18.8	7.1

covariates. This loss of information can be measured by the so-called fraction of missing information (FMI), which can be estimated via

$$\widehat{FMI} = \frac{B_M}{\bar{U}_M + (1 + M^{-1})B_M}, \tag{6}$$

where B_M is the between-imputation variance and $\bar{U}_M$ is the within-imputation variance. Specifically, let $(Y^{(m)}, U^{(m)})$ be the point and variance estimate of the service rate based on the m-th imputation. The between and within imputation variances are estimated respectively by

Table 7. Percentage rates of Asian last year service use.
New: *observed new design rates*. Imp: *imputed old design rates*. Old: *observed old design rates*

	Vietnamese			Filipino			Chinese			Other Asian		
	New	Imp	Old	New	Imp	Old	New	Imp	Old	New	Imp	Old
Specialist (1,4,7,11)	3.6	5.7	3.5	3.4	4.6	1.8	6.2	5.6	3.2	2.0	5.8	2.5
1. Psychiatrist	3.6	4.6	3.3	1.4	1.8	0.9	4.8	2.6	1.4	0.6	4.0	1.5
4. Psychologist	2.5	1.1	0.2	3.4	2.4	0.4	4.2	3.3	1.8	1.1	2.1	0.9
7. Other M.H. Prof.	2.5	1.1	0.4	1.4	1.3	0.5	3.9	0.5	0.0	0.9	1.1	0.2
11. Hotline	0.0	0.4	0.0	1.4	0.8	0.3	0.0	0.9	0.7	0.4	0.4	0.0
Generalist (2,3,8)	16.4	9.2	3.2	12.8	10.7	2.8	5.8	5.1	1.9	5.3	8.0	3.7
2,3. Other Med. Doc.	16.4	9.1	3.2	12.3	9.9	2.3	5.8	5.0	1.9	5.3	7.5	3.7
8. Other Professionals.	2.5	0.5	0.0	4.1	1.5	0.8	2.5	0.6	0.0	0.6	1.0	0.0
Human Services (5,6,9)	1.2	2.9	1.0	5.5	5.8	1.5	4.1	3.9	1.4	5.8	4.1	1.1
5. Social Worker	0.0	0.4	0.0	2.3	2.0	0.9	2.6	0.4	0.0	3.3	1.3	0.9
6. Counselor	0.7	1.7	0.7	2.9	2.8	0.8	0.6	2.0	0.8	0.0	1.9	0.3
9. Religious Advisor	1.2	1.2	0.3	5.0	2.0	0.1	2.9	2.1	0.8	2.4	2.3	0.1
Alt Services (10,12.13)	2.8	2.9	1.0	3.1	3.9	0.9	1.4	3.3	1.9	3.2	5.4	2.7
10. Other Services	2.4	1.8	0.4	1.3	1.5	0.3	0.4	1.0	0.2	2.7	3.3	1.3
12. Internet	0.5	0.9	0.2	1.5	1.7	0.7	1.0	1.6	1.1	1.4	2.2	1.6
13. Self Services	0.0	0.8	0.5	2.7	1.4	0.2	0.0	0.9	0.5	0.0	0.8	0.3
Formal Services (1–8)	17.9	12.4	6.2	15.4	15.2	5.0	8.5	8.3	4.5	9.5	10.8	5.1
Any Services (1–10)	20.7	13.6	6.4	16.8	16.8	5.1	9.6	9.9	5.0	12.8	13.2	6.4
Any Services (1–13)	21.1	14.2	6.5	16.8	17.9	5.7	9.9	11.4	6.6	13.4	14.2	7.8

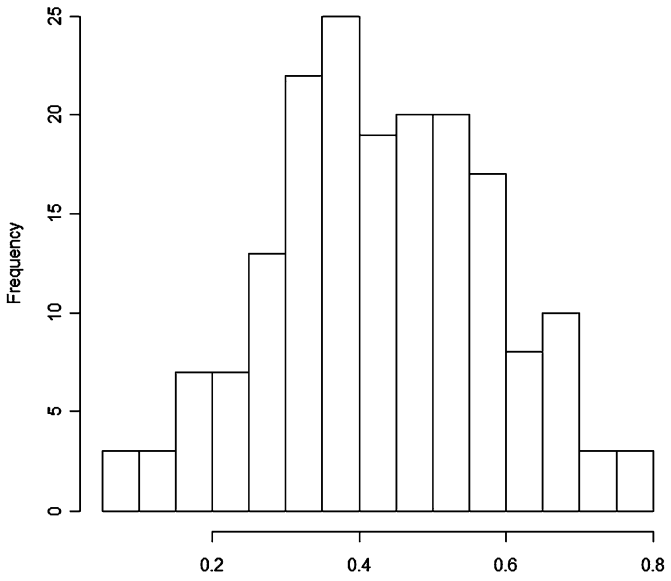

Figure 2. Estimated fraction of missing information.

(7)
$$B_M = \frac{1}{M-1}\sum_{m=1}^{M}(Y^{(m)}-\bar{Y}_M)^2 \quad \text{and} \quad \bar{U}_M = \frac{1}{M}\sum_{m=1}^{M}U^{(m)},$$

where $\bar{Y}_M$ is the average of $\{Y^{(m)}, m = 1, \ldots, M\}$; see [22].

Figure 2 shows the histograms of 180 FMIs, computed for the 20 services in Table 4 for nine ethnic groups: four Latino groups, four Asian groups, and one white group. The median of the estimated FMI is 43%, the 25% and 75% quantiles are 33% and 53% respectively, which are much less than 75%, indicating that, for the majority of the variables, there is a gain by imputation. We do notice, however, that the maximal FMI is 80%, which seems odd as it exceeds the 75% maximal loss of information. However, we must be mindful that the FMI measure here is based on an asymptotic normality assumption and it is subject to estimation errors and Monte Carlo errors. Therefore, having a few FMI that slightly exceed the 75% limit actually is an indication that the FMIs estimates given here are realistic, especially as our estimates are not numerically constrained in any way other than being positive and bounded above by $M/(M+1)$, a universal factor due to the finite number of imputations M; see [22].

3.2 A diagnostic statistic

Once multiple imputations are created, one natural question is how do we know if they are good, or even what the meaning of "good" here is. This question is generally hard to answer because usually one does not have a "Gold Standard" to check against. In our current context, however, we do have the observed rates from the new-design group as the benchmark. We then obviously do not want the new design rates to be very different from the imputed rates. If that happens, we may suspect that there is a serious failure of the imputation model, or errors in the computation (e.g., the MCMC failed to converge), or some other problems.

Given the challenges, we made an attempt to construct a building block. For a given sub-population, let $\bar{Z}$ be the observed new design rate of a particular service and $\bar{Y}_M$ be the

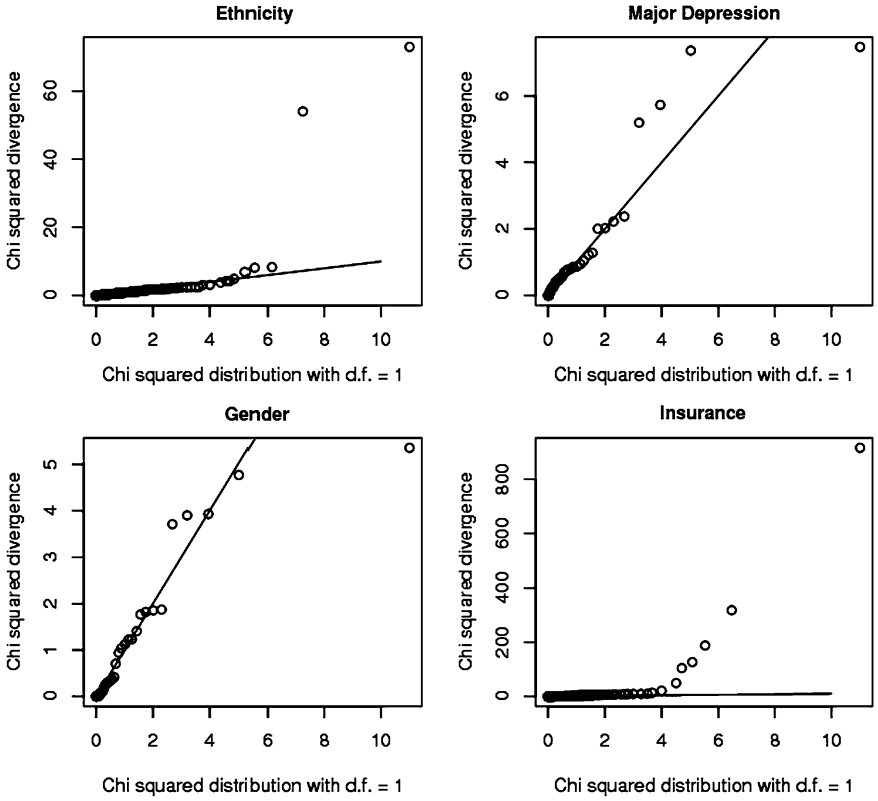

Figure 3. QQ plots of the statistic T_i against the χ^2_1 distribution with no extreme values removed.

imputed service rate of the old group, $\bar{Y}_M = \frac{1}{M}\sum_{m=1}^{M} Y^{(m)}$, where $Y^{(m)}$ is the m-th imputed rate. We construct our diagnostic statistic based on the usual χ^2 like quantity,

$$D = \frac{(\bar{Z} - \bar{Y}_M)^2}{Var(\bar{Z} - \bar{Y}_M)}. \tag{8}$$

The central difficulty here is the estimation of $Var(\bar{Z} - \bar{Y}_M)$, because it needs to take into account *both sampling variability and imputation uncertainty.* Technically, the most challenging part is to account for the strong dependence of $\bar{Y}_M$ on $\bar{Z}$, especially because of the complexity in both the imputation model and the survey design. Currently, we find it is only feasible and practical to provide an estimate of a *lower bound* on $Var(\bar{Z} - \bar{Y}_M)$.

Underestimating variances is typically unacceptable in statistical analysis and scientific investigations. However, in the current setting where a central goal is to flag problematic imputation results for further investigations, we would rather err on the side of false alarm than miss the real signal, when we cannot do it correctly. Furthermore, the proposed procedure is built upon somewhat heuristic and approximate arguments (see Section C in the supplemental material) and therefore the conservative nature of the proposed screening procedure, in the sense of preferring over-flagging, may help to guard against errors in approximations that head in the direction of under-flagging.

To proceed, let S be the service variable whose rate is being checked, and $\mathcal{H}$ denote all the covariates and the remaining 12 service variables (not counting those aggregated "any service" variables). Then, as will be argued in Section C in the supplemental material, asymptotically,

$$Var(\bar{Z} - \bar{Y}_M) \geq \frac{V^{new}}{n^{new}} + \frac{V^{old}}{n^{old}} + \frac{E(B_M)}{M}. \tag{9}$$

Here B_M is given in (7), n^{old} and n^{new} are the effective sample sizes (which will be estimated via (8) in Section A in the supplemental material) for the old and new groups respectively, and

$$V^{new} = Var(E(S^{new}|\mathcal{H})), \qquad V^{old} = Var(E(S^{old}|\mathcal{H}))$$

are the sampling variances of the conditional expectations of the (single) service variable given $\mathcal{H}$, where the superscript "new" and "old" indicate which design group. We need to treat V^{new} and V^{old} separately because the service variables are not completely observed in the old group due to underreporting. Therefore, the information contained in $\mathcal{H}$ is more for the service variable under the new design than under the old design. Indeed, this loss of information causes further

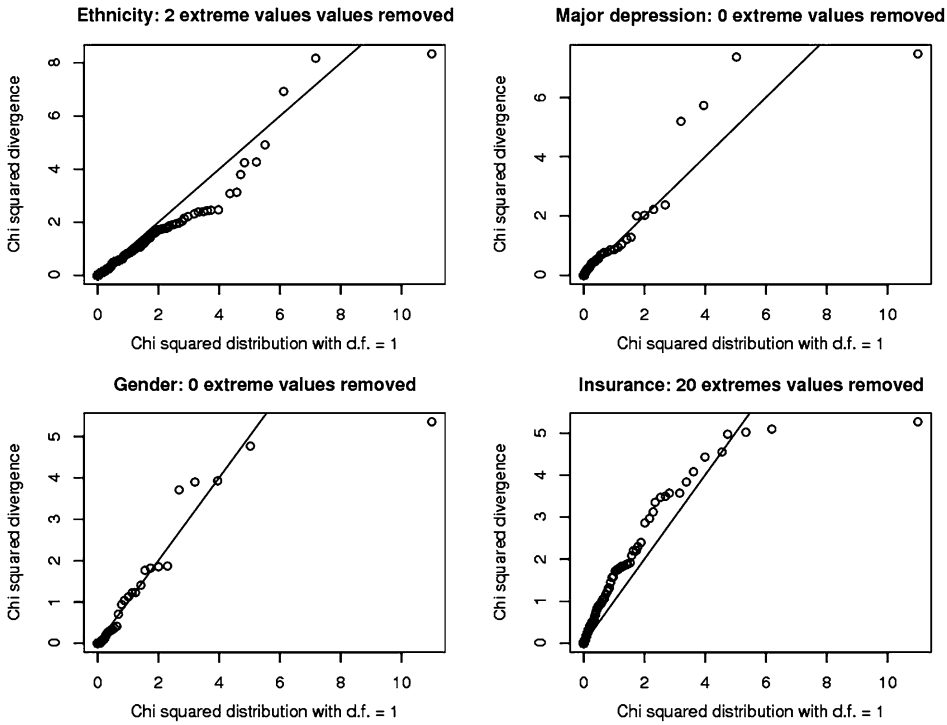

Figure 4. QQ plots of the statistic T_i against χ^2_1 distribution after the extreme values removed.

trouble in estimating V^{old}, and therefore we will again have to use an estimate of a lower bound on V^{old} in forming an estimator of a lower bound of $Var(\bar{Z} - \bar{Y}_M)$:

$$\hat{V}_{low} = \frac{\hat{V}^{new}}{n^{new}} + \frac{\hat{V}^{old}_{low}}{n^{old}} + \frac{B_M}{M},$$

where the subscript "low" indicates that a lower bound is used. The expressions of $\hat{V}^{new}$ and $\hat{V}^{old}_{low}$ are given in Section C in the supplemental material, under the assumption that our imputation model is adequate, which can be treated as the null hypothesis for setting up our diagnostic test statistic:

$$\tilde{D} = \frac{(\bar{Z} - \bar{Y}_M)^2}{\hat{V}_{low}}.$$

Let $D_1, \ldots, D_N$ be the diagnostic statistics for different strata and services. For instance, if we stratify the population by gender, then $N = 52$ (26 services by 2 strata). See the figures in Section B in the supplemental material for graphical comparisons. If the quality of imputation is acceptable, the empirical distribution of D_i's is expected to resemble that of a χ^2_1 distribution. We assess departures from this expectation by the following procedure. We remove a certain number (as few as possible) of largest D_i's such that the distribution of the remaining D_i's is close to or stochastically dominated by the distribution of χ^2_1. We denote this number by η. A large value of η will raise the warning flag, suggesting further investigations.

Figure 3 shows a number of Q–Q plots of the D_i's against χ^2_1. The Q–Q plots for the stratification by major depression and gender lie approximately on the 45 degree line. But the Q–Q plot for ethnicity shows two extremely large values, while that for insurance has quite a few very large values. Figure 4 shows the Q–Q plot of the D_i's against χ^2_1 after removing the extremely large values. For ethnicity, after removing two extreme values in the Vietnamese group, the distribution of the rest D_i's becomes reasonably close to that of χ^2_1, indicating the problem lies in the Vietnamese group, as we noticed before. For insurance, we had to remove the 20 largest D_i's before the distribution looks approximately like χ^2_1. Out of the 20 extreme values, 16 are in the "other insurance" stratum, which is known to be problematic (see Section B in the supplemental material). This demonstrates that our screening procedure is doing a decent job in flagging trouble spots for further investigation.

3.3 Self-criticism

Of course, our screening procedure is far from perfect, so is our imputation model despite literally years of effort devoted to this project. Much more needs and can be done for developing both of them, but we simply had to complete the project and provide the "deliverable" as a part of our funding requirements. This case study therefore reminds us extremely well the joy and frustration of doing applied statistics, and most importantly the need for developing and

teaching statistical techniques that take into account time and resource constraints in principled ways.

ACKNOWLEDGEMENT

We appreciate the editors and the reviewer for providing valuable comments. This research is supported in part by NSF and NIH.

Received 30 December 2012

REFERENCES

[1] Agresti, A. *Categorical Data Analysis*. Wiley, NY, 1990. MR1044993

[2] Alegria, M., Takeuchi, D., Canino, G., Duan, N., Shrout, P., Meng, X. L., Vega, W., Zane, N., Vila, D., Woo, M., Vera, M., Guarnaccia, P., Aguilar-Gaxiola, S., Sue, S., Escobar, J., Lin, K. M., and Gong, F. Considering context, place and culture: the national latino and asian american study. *International Journal of Methods in Psychiatric Research*, 13(4):208–220, 2004.

[3] Anderson, T. W. *An Introduction to Multivariate Statistical Analysis*. Wiley-Interscience, 3rd edition.

[4] Barnard, J. and Rubin, D. B. Miscellanea. Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4):948–955, 1999. MR1741991

[5] Clayton, D. G. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151, 1978. MR0501698

[6] Cox, C. Multinomial regression models based on continuation ratios. *Statistics in Medicine*, 7:435–441, 1988.

[7] Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–220, 1972. MR0341758

[8] Duan, N., Alegria, M., Canino, G., McGuire, T. G., and Takeuchi, D. Survey conditioning in self-reported mental health service use: Randomized comparison of alternative instrument formats. *Health Services Research*, 42(2):890–907, 2007.

[9] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. *Bayesian Data Analysis*. CRC Press, 2nd edition. MR2027492

[10] Gelman, A. and Meng, X. L. *Applied Bayesian Modeling and Causal Inference from Incomplete Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*. Wiley, John & Sons, Incorporated, 2004. MR2134796

[11] Heagerty, P. J. and Zeger, S. L. Multivariate continuation ratio models: Connections and caveats. *Biometrics*, 56(3):719–732, 2000. MR1791148

[12] Heeringa, S., Wagner, J., Torres, M., Duan, N., Adams, T., and Berglund, P. Sample designs and sampling methods for the collaborative psychiatric epidemiology studies (cpes). *International Journal of Methods in Psychiatric Research*, 13:221–240, 2004.

[13] Li, K. H., Meng, X. L., Raghunathan, T. E., and Rubin, D. B. Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica*, 1:65–92, 1991. MR1101316

[14] Liu, J. C. *Effective modeling and scientific computation with applications to health study, astronomy, and queueing network*. PhD thesis, Harvard University, Cambridge, MA, May 2008. MR2711683

[15] Liu, J. C., Meng, X. L., Alegria, M., and Chen, C. Multiple imputation for response biases in nlaas due to survey instruments. *ASA Proceedings of the Joint Statistical Meetings*, pages 3360–3366, 2006.

[16] McCullagh, P. and Nelder, J. A. *Generalized Linear Models, Second Edition (Monographs on Statistics and Applied Probability)*. Chapman & Hall, 1989. MR0727836

[17] Meng, X. L. Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, 9:538–573, 1994.

[18] Meng, X. L. and Rubin, D. B. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993. MR1243503

[19] Oakes, D. Bivariate survival models induced by frailties. *Journal of the American Statistical Association*, 84:487–493, 1989. MR1010337

[20] Prentice, R. L. and Gloeckler, L. A. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, 34:57–67, 1978.

[21] Robins, L. N., Helzer, J. E., Croughan, J. L., and Ratcliff, K. S. National institute of mental health diagnostic interview schedule: its history, characteristics and validity. *Archives of General Psychiatry*, 38:328–366, 1981.

[22] Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys*. John Wiley, NY. MR0899519

[23] Shapiro, G. M. Interviewer-respondent bias resulting from adding supplemental questions. *Journal of Official Statistics*, 3:155–168, 1987.

[24] van Dyk, D. A. and Meng, X. L. The art of data augmentation (with discussion). *Journal of Computational and Graphical Statistics*, 10(1):1–111, 2001. MR1936358

Jingchen Liu
1255 Amsterdam Ave
Department of Statistics
New York, NY 10027
USA
E-mail address: jcliu@stat.columbia.edu

Xiao-Li Meng
1 Oxford Street
Department of Statistics
Cambridge, MA 02138
USA
E-mail address: meng@stat.harvard.edu

Chih-nan Chen
National Taipei University
67, Sec. 3, Ming-shen E. Rd.
Taipei, 10478 Taiwan
Taiwan
E-mail address: cnchen@mail.ntpu.edu.tw

Margarita Alegria
120 Beacon Street
4th Floor
Somerville, MA 02143
USA
E-mail address: MAlegria@charesearch.org